

Caldwell, C. A. & Millen, A. E. (2010). Human cumulative culture in the laboratory: effects of (micro) population size. *Learning and Behavior*, 38, 310-318 (doi:10.3758/LB.38.3.310).

Human Cumulative Culture in the Laboratory: Effects of (Micro) Population Size

Christine A. Caldwell

Ailsa E. Millen

Department of Psychology, University of Stirling, UK

Traditionally, experiments on social learning (both in humans and nonhumans) involve dyads, with an experimenter or experimenter-trained conspecific serving as the demonstrator and the participant as the observer. But social learning in nature often involves multiple potential models, and the models themselves were once learners. We discuss our studies of social learning in adult humans in interactive group settings in the absence of formal demonstrations by experimenters, which track transmission over multiple learner generations. In these experiments we find evidence for cumulative learning over generations. This has allowed us to manipulate learning conditions in order to test hypotheses regarding the necessary conditions for cumulative culture. We also report results from a further experiment using similar methods, which compares conditions of varying cohort size. Participants were given the task to build a paper airplane to fly as far as possible. Contrary to expectations, there was no advantage for larger cohort sizes, in terms of the cumulative effects observed.

Introduction

Capturing Social Learning in Natural Contexts. In much of the experimental literature on social learning, the ecological validity of the social context is often sacrificed in the interests of greater control over the material to which the participants are exposed. It is by no means a criticism to draw attention to the tight control involved in these designs, and indeed such approaches have proven to be highly successful in recent years, resulting in important advances in both the nonhuman and human social learning literature.

For example, using such methods, Horner and Whiten (2005) were able to show intriguing differences between chimpanzees (*Pan troglodytes*) and human children in terms of the elements of a demonstration which were copied. Following a demonstration by an adult human experimenter, the chimpanzees omitted actions that appeared not to be causally related to the goal (of extracting a reward from the experimental apparatus). In contrast, young children aged three and four years old copied all elements of the demonstration, irrespective of how relevant to the goal they appeared to be. Whiten, Horner, Litchfield and Marshall-Pescini (2004) review numerous studies of social learning in nonhuman primates using similar methods. Children's language learning has also been studied using similar dyadic experimenter-subject designs. Experiments show that young children can readily incorporate novel nonsense words into their vocabulary, and use these appropriately, after hearing such words spoken by an adult experimenter (for a review see Tomasello, 2000). The effects of social influences on adult humans have also been studied in a similar fashion, with manipulations achieved through the actions of experimental confederates. Experiments by Sherif (1936) and Asch (1955) show that participants asked to make a simple perceptual judgment would alter their responses if these were

not consistent with responses given by actors posing as fellow participants in the same experiment. In a more recent study following the same basic logic, Eriksson and Coultas (2009) found that participants' responses to opinion questionnaire items could be influenced by the apparent responses of previous participants (which were actually faked in order to directly manipulate the magnitude of the apparent majority).

As these studies illustrate, controlled manipulations of modeled behaviors can reveal much about what can be transmitted. And yet, when we look at real cultures and traditions, i.e. those that occur spontaneously in natural populations, there is an inevitable shift of focus due to the kind of data involved. Different issues arise as the important questions, and a rather different perspective tends to be taken, with populations, rather than individuals, representing the principal units of study.

For example, research on traditions in nonhuman primates began with reports of novel behaviors spreading through populations. Early reports of this kind were given for Japanese macaques, *Macaca fuscata*, concerning a number of new behaviors including food-washing, bathing, and acceptance of particular unfamiliar foods into the diet (see Itani and Nishimura, 1973; Kawai, 1965; Kawamura, 1959). The researchers report data on the rate at which these behaviors were adopted within the populations concerned, and also on the relationships between individuals who appeared to have influenced each others' adoption of the practice in question. More recently, studies of white-faced capuchin monkeys, *Cebus capucinus*, in the field have documented traditions in both their social (Perry et al., 2003) and foraging (Panger et al., 2003) behaviors. These studies have documented both the emergence and extinction of particular behaviors (Perry et al., 2003), and the social relationships between individuals that share particular forms of a behavior (Panger et al., 2002). See Reader and Biro (this issue) for a comprehensive review of social learning experiments in wild animals.

In humans, data on real trends in human populations has been used to test predictions about the dynamics of transmission. Bentley and colleagues (Bentley, Hahn & Shennan, 2004; Bentley, Lipo, Herzog & Hahn, 2007) have used data from popularity charts of baby names, music sales, and dog breeds, finding that the steady rates of turnover observed in each are consistent with random copying. The implication of this is that changes in the frequencies of such cultural traits over time will be determined purely by random events (analogous to neutral genetic drift). Interestingly, Henrich (2001) came to the opposite conclusion when considering the adoption curves associated with the diffusion of novel inventions and ideas (from Rogers, 1995). Henrich (2001) demonstrated that the typical S-shaped adoption curves are likely to be attributable to biased social transmission processes such as conformist learning, and appear inconsistent with unbiased transmission (although see Hoppitt, Boogert & Laland, this issue, for discussion of the diagnostic use of diffusion curves). These apparently contradictory conclusions highlight an important issue: the particular datasets chosen may impact greatly on the eventual conclusions. It may be the case that unbiased transmission is more common in certain contexts. Alternatively it could be the case that a population-level mix including both conformist and anti-conformist learning (or indeed forms of frequency dependent copying which do not quite exhibit the disproportionate influence of strong conformity) could generate outcomes that are impossible to differentiate from random copying in popularity chart data (Mesoudi & Lycett, 2009).

Studies of social transmission in natural human languages have likewise predominantly addressed questions about rates of change over time, and the emergence of new forms out of old (Lieberman, Michel, Jackson, Tang & Nowak

2007; Pagel, 2009; Pagel, Atkinson & Meade, 2007). Analyses of the similarities and differences between languages have allowed linguists to generate family trees, which indicate how closely related particular languages are in terms of common cultural descent (Pagel, 2009).

Studies of real traditions therefore tend to concern not the precise nature of what can be transmitted, but shifting trends across the population, i.e. the patterns with which particular preferences or skills catch on, or die out. Dyadic experiments involving a pairing of experimenter or trained demonstrator and participant simply cannot capture the dynamic, evolutionary properties of social learning (e.g. Mesoudi, Whiten & Laland, 2006) since they only involve transmission over one “generation”. Furthermore, it is clear when considering studies of real traditions that, in natural contexts, social learning involves multiple influences on behavior, and that there is rarely a clear-cut answer to the question of whom exactly a particular trait might have been copied from. Dyadic social learning experiments also fail to capture this feature of social learning, since there is only one potential model.

Transmission Chain Designs. There are, however, some experimental designs which are specifically devised in order to study the effects of repeated transmission over multiple generations. Mesoudi et al. (2006) have described such designs as using the “transmission chain method” (p339). Essentially in these experiments, naïve participants are initially exposed to some novel stimulus material, or a demonstrated action, as in typical dyadic social learning experiments. However, the key difference is that these participants then go on to take the role of demonstrator, in that their responses then form the stimulus material for the subsequent participant. In this way, researchers can study how information can be shaped, degraded, or built upon, over repeated transmission events. Since these studies give an indication of the longevity of particular behaviors over repeated transmission, they provide a much better insight into the sorts of traditions that might actually be sustained in real populations.

An example of one such study was the pioneering work of Jacobs and Campbell (1961), following up on the studies by Asch (e.g. 1955) and Sherif (1936), mentioned earlier. Whilst it was well known that participants could be induced to give responses which were in direct conflict with their perceptual experience (as shown by Sherif and Asch), Jacobs and Campbell (1961) were interested in the extent to which such counter-intuitive beliefs could be perpetuated through a group by social transmission alone. Jacobs and Campbell (1961) established experimental groups which were founded by confederates, instructed to respond with a significant overestimation of their true perception of the strength of a visual movement illusion. However, the confederates were gradually replaced by naïve participants. In an example of one of their conditions, there were three individuals present in the group at any one time, and at the start of the experiment two of these individuals were confederates and one was a naïve participant. Each individual from the group was asked to estimate the degree of the illusory movement perceived, starting with the confederates, and their responses were recorded. This was repeated thirty times, after which one of the confederates was removed and replaced by another naïve participant. Thirty more trials were then carried out with this new group, and the remaining confederate was then removed and replaced with a further naïve participant. This procedure continued for a total of ten “generations”. Jacobs and Campbell (1961) found that the overestimation bias induced by the confederates persisted for several generations after the final confederate had been removed. Nonetheless, in all cases, estimates of motion did eventually return to the same levels as those given by control groups which had never had any contact with confederates.

Along a similar vein, Flynn (2008) has followed up on the research by Horner and Whiten (2005), also mentioned earlier, which indicated that young children would imitate all elements of a demonstrated action, including those that were irrelevant to the goal of extracting a reward (see also Flynn, this issue). Flynn (2008) used the same task as Horner and Whiten (2005), and tested a similar age group. However, she used a transmission chain design in which the child participant became the demonstrator for the subsequent participant. In this way she was able to study the level of fidelity with which the behavior was transmitted in eight chains, each composed of five or six children. Interestingly, the irrelevant actions dropped out of the demonstrations relatively rapidly, with most chains converging on the most efficient possible solution by the final participant. The implications of this study are therefore very much in line with those of Jacobs and Campbell's (1961) experiment: although arbitrary and functionless behaviors and beliefs can certainly be transmitted, at least over a single generation, repeated transmission will often filter these out. It should be noted however, that Flynn's (2008) study necessitated that young children played the role of demonstrator from the second generation onwards, and that children may have been less likely to "overimitate" the actions of their peers, compared with those of an adult. In contrast, Horner and Whiten's (2004) study used only adult demonstrators.

Designs involving repeated transmission have also been used with nonhuman species, also generating valuable insights (see Whiten & Mesoudi, 2008, for a review of such studies). An early example was Menzel, Davenport and Rogers's (1972) study of chimpanzees' responses to novel objects. Menzel et al. (1972) made use of what they described as a "serial-grouping paradigm" (p162), which was in fact very similar to the design described above for Jacobs and Campbell's (1961) study. The sample tested consisted of 19 captive chimpanzees with limited experience of object interaction, and the researchers studied them as 17 successive groupings of three animals. These trios lived together for at least two months, at which point a new trio was formed by removing the longest-standing member, and replacing them with a naïve individual. Menzel et al. (1972) found that the first and second trios avoided both of the test objects, but one member of the third trio habituated to one object and began interacting with it, and one member of the fifth trio likewise habituated to the other test object. Subsequent trios would then interact with these objects, continuing the "play" tradition, even once the originator had left the group. Menzel et al.'s (1972) results imply that for these chimpanzees, the cultural history of the group's interaction with the objects was a better predictor of how any given individual would react to them, rather than that individual's own temperament. Again, this is a result which could not be obtained using a simple dyadic design.

Within the literature on human language learning, transmission chain designs are proving to be particularly informative. It has been suggested that certain features of language may actually be best explained in terms of the filtering process generated by repeated transmission over multiple generations of learners (e.g. see Christiansen & Chater, 2008; Kirby, Dowman & Griffiths, 2007). Kirby and colleagues (e.g. Kirby et al., 2007; Kirby & Hurford, 2002) have used this idea to develop an "Iterated Learning Model" of language structure, with iterated learning referring to "a process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way." (Kirby, Cornish & Smith, 2008, p10681). Kirby et al. (2008) have thus made use of an experimental chain design in order to examine whether this emergence of structure could be studied on a small scale in the laboratory. Kirby et al. (2008) exposed participants to an initially

structureless artificial language, by pairing strings of labels with particular visual stimuli. Participants were then tested on their knowledge of the language by asking them to produce the appropriate string of labels when shown further images. The strings produced during testing, plus the corresponding visual stimuli, were used as training materials for the next participant. It was found that, over ten generations of learners, the languages evolved to become more structured and easier to learn.

Laboratory Models of Cumulative Culture in Adult Humans. In our own work (e.g. Caldwell & Millen, 2008a; 2008b; 2009, 2010) the repeated transmission (or iterated learning) that is intrinsic to culture, whether human or nonhuman, is similarly fundamental to the phenomenon we wish to investigate. Repeated transmission in humans often results in an accumulation of inventions and technologies as later generations build on the skills and knowledge of previous ones (e.g. Boyd & Richerson, 1985; Henrich & McElreath, 2003). Tomasello (e.g. 1999) has described this cumulative, transgenerational learning as a “ratchet effect” (p5), to capture the way in which practices tend to be faithfully preserved in the population until some improvement or elaboration is invented. Capturing this aspect of human social learning in the laboratory requires an approach which goes beyond dyadic designs.

In our research we have also used methods which involve repeated transmission, which involve no formal demonstrations at all, even at the very beginning of the chain. Caldwell and Millen (2008a) ran twenty chains, each composed of ten participants. Participants were presented with simple technical challenges, which each had clear objective measures of “success” which could be used to evaluate the performance of participants over generations. As in the studies by Jacobs and Campbell (1961), and Menzel et al. (1972), described previously, the complete chain of participants were separated into small test groups who were together concurrently, with the longest-standing member of the group removed, and replaced by a naïve individual, at regular intervals until all members of the chain had taken part. Figure 1 (“Two Model Condition”) shows how the membership of the test group changed over a complete trial in Caldwell and Millen (2008a). Following Jacobs and Campbell (1961) and Baum, Richerson, Efferson & Paciotti (2004) we have referred to this type of design as a “microculture” (Jacobs & Campbell, 1961), or “microsociety” (Baum et al., 2004), capturing the notion of the chain as a population in miniature, with concurrent participants representing a cohort of contemporaries.

Caldwell and Millen (2008a) gave ten chains of participants the task of building a paper airplane to fly as far as possible, and gave ten chains the task of building a tower from spaghetti and modeling clay, to be as high as possible. The respective goal measures (of flight distance and tower height) could be used to evaluate performance on the task. Consistent with the cumulative learning expected, in both cases later participants were found to perform better on these tasks than earlier ones. The real value of these methods has been to open up the possibility of testing hypotheses concerning the necessary conditions for cumulative culture. Although real human traditions can provide rich data indicating that cumulative cultural evolution has occurred, the microsociety method allows us to actively manipulate particular variables that may influence whether or not cumulative effects are observed. A great deal of attention has been paid to the phenomenon of human cumulative culture in the recent research literature, and there has been much speculation about the possible reasons for its apparent absence in nonhuman traditions (see Caldwell & Millen, 2008b, for a review of the key perspectives). Our laboratory models allow some of these theories to be put to the test.

Attempting to do just that, Caldwell and Millen (2009) used the paper airplane task in order to compare conditions in which participants had access to different social learning mechanisms, separating out information in the form of actions (allowing for imitation), results (allowing for emulation), and teaching. Despite assertions that imitation and teaching may be necessary for cumulative culture to occur (e.g. Tomasello, 1999; Boyd & Richerson, 1996), Caldwell and Millen (2009) found evidence of cumulative effects even when participants only had access to information in the form of results, and were therefore restricted to emulative learning (from finished products).

Population Size and Cumulative Culture. The aim of the current experiment was to test hypotheses concerning population size and cumulative culture. As mentioned previously, dyadic social learning experiments not only neglect the repeated transmission involved in real traditional behavior, but also fail to capture the fact that any given learner may have access to multiple potential models. This is also an important feature of social learning in natural contexts, and researchers have devoted considerable attention to the adaptive strategies that learners could exploit in order to choose between observed behavioral alternatives (e.g. Boyd & Richerson, 1985; Kendal, Coolen, van Bergen & Laland, 2005; Laland, 2004).

The matters of population size and model selection are particularly pertinent to the issue of cumulative culture. Henrich (2004) has developed a formal model of cumulative cultural evolution which illustrates the importance of population size when learners are attempting to maximize their own skill level by copying the most skilled model available. In Henrich's (2004) model, members of a population attempt to learn from the most skilled individual, but learning is imperfect such that transmission typically leads to some loss of skill. Henrich (2004) has shown that when population size increases, the effects of model selectivity (which becomes more effective the more models one has to choose from) and occasional serendipitous copying errors (which likewise increase in number with increasing population size) will outweigh the losses involved in most cases of transmission. The net effect will therefore be an increasing skill level in the population. Henrich (2004) has used this model to illustrate how it would be possible for a small population to experience maladaptive skill loss, and posits this as an explanation for the apparent decline in the variety and complexity of material culture from the population of Tasmania between 8000 B.P. and 3000 B.P.

Powell, Shennan and Thomas (2009) have extended Henrich's model, taking into account subpopulation density, and migration between subpopulations, as well as total population size. Powell et al. (2009) use the results of simulations to argue that the timing of the Upper Paleolithic transition (involving a marked escalation in technological and cultural complexity) is likely attributable to demographic, rather than cognitive, changes.

Using our experimental methods, it is relatively easy to manipulate this variable of population size. Caldwell and Millen (2008a) used test group cohorts consisting of four participants, two of whom were building, and two of whom were observing (see Figure 1, "Two Model Condition"). In the current experiment therefore, we simply manipulated this variable of cohort size, such that participants either had access to a greater (Figure 1, "Three Model Condition"), or lesser number of potential models (Figure 1, "One Model Condition"). Using our paper airplane task described previously, it was predicted that access to a greater number of models would result in greater increases in skill level in the population (as measured by flight distance).

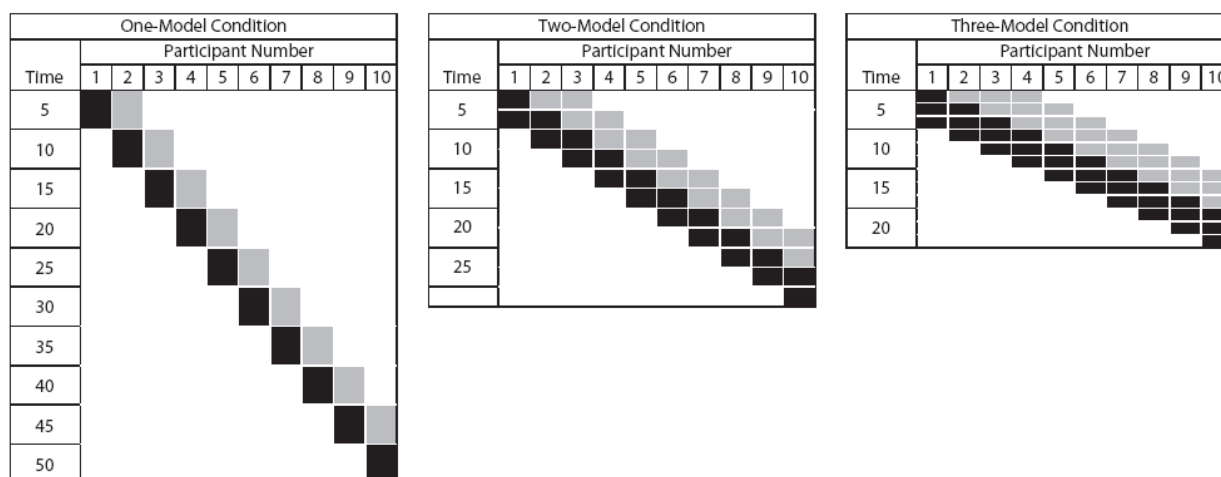


Figure 1. Membership of the test group at any one time during a trial, for the three conditions. Shading indicates the role of the participant: Black indicates building, gray indicates observing. Times are given in minutes.

Method

Participants. Participants were recruited on campus at the University of Stirling and the University of Edinburgh. Three hundred participants took part (ten chains of ten individuals for each of the three conditions). Data for the two model condition (100 participants) were collected as part of a previous experiment (reported in Caldwell & Millen, 2009, as condition AR). The remaining 200 participants (in the one model and three model conditions) were recruited specifically for the purpose of making the comparison of the effects of test group size.

Participants took part in exchange for a participation fee, paid at a rate of £5 per hour. University of Stirling Psychology students were offered the choice between the participation fee and a research participation credit. All participants (whether given the credit or the fee) were also given a monetary performance incentive to encourage them to score as well on the task as possible (an extra ten pence for each meter of flight achieved, which resulted in participants receiving an average of an extra 60 pence).

The mean age of the participants was 21.40 years ($SD = 5.98$, youngest = 16, eldest = 52), and 68% of the sample were female (totals = 95 males, 205 females). Ethical approval for this research was granted by the University of Stirling Department of Psychology Ethics Committee. The procedure was explained to all participants in advance, and they each gave written consent to participation.

Procedure. For each trial the full chain of ten participants was assembled prior to the experiment commencing. They were informed that they were about to take part in a team challenge, and that they would be called in turn to engage in the task. Participants were each randomly assigned a number between one and ten to indicate their position in the chain. They completed consent forms and then waited their turn to join the group in the test area. The test area could not be seen by the waiting participants.

When participants joined the group in the testing area they were provided with written instructions about the nature of the task. They were informed of the aim of the task (i.e. to build a paper airplane that flew as far as they could make it go) and of their time restrictions (five minutes of observation time followed by five minutes in

which to build their own plane). They were also informed that their participation fee would be increased in proportion with their performance on the task (as detailed in Participants section above). Figure 1 shows schematics of the membership of the test groups for the three different conditions, indicating the role of each participant at any given time during a single trial. Participants were encouraged to observe and learn from others, but were expressly forbidden from building another individual's plane for them. Thus they were instructed not to touch another individual's piece of paper. Each participant was provided with a single sheet of A4 paper with which to create their airplane.

Within the test group, participants were kept aware of their current role (observing or building) and the time elapsed, by a computer display and reminders from the experimenter. Once an individual's five minute building period was up, their plane was evaluated. This involved the participant throwing their plane three times, with the experimenter recording the distances flown, and then taking the best of the three measurements (to allow for mis-throws). Participants in the test group could see planes being thrown. Once a plane's flight distance had been recorded, it was then placed on display in the test group, and the experimenter wrote down the measurements next to each, so that this information was also available. Each plane was held on display for five minutes, when it was removed and replaced by the next most recently completed plane.

As can be seen from Figure 1, the total time of a complete trial varied across the experimental conditions. It was important to ensure that in all conditions, each participant had the same amount of time in which to build their plane. The manipulation of test group size therefore meant that the three model condition progressed more rapidly than the two model condition, and the two model condition more rapidly than the one model condition. It should be noted that the number of completed planes on display at any one time also varied across conditions in the same way as the number of models in the test group. As planes were always placed on display for five minutes after completion, this meant that in the one model condition there was only ever one plane on display at any one time, whereas in the two model condition there were two, and in three model condition three.

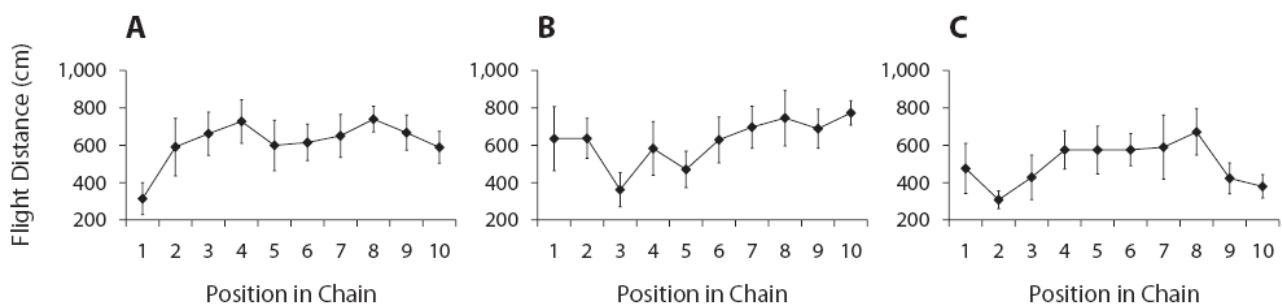


Figure 2. Mean flight distances of the paper airplanes produced by the participants in Positions 1–10 of the chains. (A) The one-model condition. (B) The two-model condition. (C) The three-model condition. Error bars indicate ± 1 SEM.

Results

It was expected that there would be evidence of cumulative learning in all three experimental conditions. Page's *L* Trend Test (Page, 1963) was used to analyze the complete data over the ten generations to explicitly test for successive increases in the flight distance score (Caldwell & Millen, 2008a; 2009). As reported in Caldwell & Millen (2009) this was significant for the two model condition ($L = 3221$, $k = 10$, $N = 10$, $p = 0.016$). Cumulative improvement was also found for participants in the one model condition ($L = 3250.5$, $k = 10$, $N = 10$, $p = 0.005$). However, the effect of improvement over generations did not reach significance for the three model condition ($L = 3117$, $k = 10$, $N = 10$, $p = 0.145$). Figure 2 displays the data for all three conditions.

In order to compare conditions, data were collapsed across the first three (positions 1, 2 and 3) middle three (4, 5 and 6), and final four (7, 8, 9 and 10) participants in each chain. Means are reported in Table 1. Tests indicated that the distribution of these values was not significantly different from the normal distribution. These values therefore allowed us to perform a 3x3 ANOVA, with generation (early, middle and late) as a repeated measures variable, and experimental condition (one, two or three model) as a between-subjects variable.

Table 1
Mean Flight Distances (With Standard Deviations)
for Participants From Early, Middle,
and Late Generations in the Chain

Condition	Early Generations (Positions 1–3)		Middle Generations (Positions 4–6)		Late Generations (Positions 7–10)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
One model	522.0	308.6	647.0	241.6	661.8	201.5
Two models	544.6	282.4	560.6	297.8	726.1	185.0
Three models	403.8	248.1	575.2	244.2	515.8	198.4

It should be noted that the choice to collapse data across these particular chain generations was partly arbitrary. The first three generations were grouped due to the fact that the effects of the group size manipulation ought only to be fully apparent from participant four onwards (being the first participant to actually experience three models in the three-model condition). However, the split between middle and late generations was made purely to generate relatively even group sizes for statistical comparison. The relevant statistics (ANOVA and t-tests) have also been run using generations 4, 5, 6 and 7 as the “middle” grouping, and generations 8, 9 and 10 as the “late” grouping. There was no change to the significance of the ANOVA results reported below. This alternative split does however result in one more post-hoc comparison and one more t-test reaching significance. The relevant comparisons are identified in the results. All other results arising from this alternative grouping are not discussed further for reasons of brevity as they are consistent with those reported below.

There was a main effect of generation [$F(2,54) = 3.345, p = 0.043$], consistent with our finding that scores were higher later in the chain. Post-hoc comparisons (Fishers LSD) indicated that the late generations' (positions 7-10) scores were higher than those of the early generations (positions 1-3) ($p = 0.026$). The late generations' scores were not significantly different from those of the middle generations (positions 4-6) however ($p = 0.459$). The middle generations' scores were also not significantly different from those of the early generations ($p = 0.082$), although it should be noted that this comparison was significant when using the alternative split mentioned above, comparing values collapsed across positions 1-3, with those collapsed across positions 4-7 (instead of 4-6).

There was no main effect of model condition [$F(2,27) = 1.457, p = 0.251$], indicating that no condition was better overall than others. There was also no interaction between generation and condition [$F(4,54) = 0.694, p = 0.599$], so the trend towards improvement later in chains was not significantly stronger in any of the conditions over others.

Independent t-tests were run using these same collapsed values for early, middle and late in the chain, in order to explore the nature of any differences between the conditions. These are reported in Table 2. The only comparison which was significant was the comparison between the two model and three model conditions for the late generations, and this was in the direction of the two model condition scores being higher for this section of the chain, contrary to expectations. Using the alternative collapsed values discussed above (comprising positions 4-7 as the "middle" generations, and 8-10 as the "late" generations), the comparison between the one model and three model condition was also significant for the late generations. Again this was in the direction of the scores from the one model condition being higher than those of the three model condition. These comparisons should be interpreted with caution due to the number made (since all would be non-significant if applying a correction for multiple comparisons such as the Bonferroni correction). All the same, it is clear that any differences between the conditions are in fact, contrary to predictions, in the direction of the three model condition participants scoring lower on the flight distance measure, compared with those in the smaller sized groups, with any such effects restricted to later in the chain of participants.

Table 2
Pairwise Comparisons Between the Conditions for Early, Middle, and Late Generations in the Chain

Condition	Early Generations (Positions 1–3)		Middle Generations (Positions 4–6)		Late Generations (Positions 7–10)	
	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
One model–two models	−0.170	.867	0.713	.485	−0.744	.467
Two models–three models	1.184	.252	−0.120	.906	2.452	.025*
One model–three models	0.944	.358	0.661	.517	1.633	.120

Note—Comparisons carried out using independent *t* tests, $df = 18$ for all comparisons. * $p < .05$.

Discussion

Results for the one model condition were very much in line with results that we have reported to date using two model designs (Caldwell & Millen, 2008a; 2009). In the one model condition we found a significant effect of cumulative improvement over generations of participants within chains, as also found in the two model condition. Furthermore there were no apparent differences between these conditions in terms of either the overall level of performance, or the extent of the cumulative effect.

This replication of our previous findings, using a different design, has interesting implications for understanding the strategies that human participants use in order to engage in cumulative learning. It had been expected that more information (in the form of a larger test group, greater number of potential models, and greater number of completed planes to inspect) would equate to better performance on this task, and potentially more pronounced cumulative effects. However, no such differences were found between the one model and two model conditions. Participants in the one model condition saw only one person building a plane, and were exposed to a total of two completed designs (one which was on display while they observed their predecessor in the chain, and the completed design of their predecessor which was placed on display at the start of their five minute building period). Clearly this information was sufficient to generate a cumulative learning effect. It may be that participants were explicitly comparing the success of the two completed designs, and preferentially copying planes which flew further, which ought to generate a cumulative learning effect. However, it remains possible that the effect would be found even if participants were only exposed to one previous design.

In our previous work (Caldwell & Millen, 2009) we have run a condition in which participants were not provided with explicit information regarding flight distances, but did see other participants building their planes. Under these circumstances, participants did not have access to direct information about the effectiveness of particular designs. Nonetheless, participants must have been able to make judicious decisions about whether or not to copy, since even in this condition, a significant cumulative effect was found (Caldwell & Millen, 2009). It is possible that participants were able to use heuristics associated with the visual properties of the finished products in order to judge what might be a good or bad design (using, for example, apparent aerodynamic properties or symmetry). They may also have been able to make a simple decision about whether or not to copy based on their knowledge of their own expertise (and therefore the likelihood of their own design being better than another participant's), possibly also taking into account the likely expertise of the previous participants (again using heuristics such as sex, or apparent building proficiency).

The implication of this for our current experiment is that human participants do not apparently need to be able to make precise comparisons of explicit payoff information in order to engage in the strategic use of social information which results in cumulative cultural learning. An even more impoverished version of our one model condition, where participants saw only one completed design, might therefore also still generate such an effect.

The results from our three model condition are rather more difficult to interpret. Although cumulative improvement was identified in both the one model and two model conditions, the three model condition showed no significant effect of increased flight distances over generations (although this seems primarily attributable to some unusually low scores towards the end of the chains in this condition). It had

been expected that those conditions with larger test groups (and therefore a greater number of potential models) would be likely to have an advantage over those with smaller groups, and that if any difference were found it would be in this direction. This would be consistent with the mathematical models of population size described in the introduction, as well as mathematical models involving manipulation of the number of “cultural parents” (Enquist, Strimling, Eriksson, Laland & Sjostrand, 2009). However, we did not find such an effect. As well as showing no significant effect of cumulative learning, the scores for participants in the three model condition were actually somewhat lower than those for the two model condition.

As noted in the results section above, the t-tests should be interpreted with some caution due to the number of comparisons made. Nonetheless, it was important to carry out these comparisons, restricted to particular stages of the chains, in order to properly compare across conditions. For the three model condition, participants 1, 2 and 3 were not in fact exposed to three models, and their test group size was smaller than for the participants later in the chain. In comparing across the conditions therefore, we are particularly interested in the performance of participants 4-10. Any differences between conditions in the scores of the first three participants would likely be a consequence of other differences between the conditions at these early points in the chain, such as the differing amounts of observation time for the initial participants (which would in fact favor the one model condition, see Figure 1). However, no differences were found between the three conditions for these early stages in the chain. Indeed, any advantage for the two model over the three model condition appeared to be entirely restricted to later (6/7-10) in the chain.

We do not believe that the results indicate that being able to make a greater number of comparisons is detrimental to people’s ability to generate cumulative culture. Clearly, greater amounts of information about potential designs and models ought to increase the chances of an individual participant making a good choice about their own design. But our results do illustrate that the strengths of micro-society designs also bring with them limitations on how far we can extrapolate the results to real cultural phenomena. Whilst micro-society designs make the laboratory study of cultural phenomena tractable, they also place constraints on the types of behavior, and the amount of information, that can be transmitted between participants.

In our largest test groups, it seems that participants may have simply been unable to make use of all of the information available in the time allocated. As will be apparent from Figure 1, in these test groups a participant was removed, and replaced by a new test group member, every one minute and forty seconds. The completed planes were likewise replaced at the same rate. This replacement rate was necessary so that we could increase the size of the test group, whilst maintaining constant individual building times across conditions (important for obvious reasons). Participants in this condition saw three other participants complete their planes prior to beginning their own build. Furthermore, for participants later in the chain, additional completed planes would have been on display during their observation period. But the rapid rotation of the various sources of social information appears to have been counterproductive, perhaps depriving participants of the time required in order to process a particular design. For example, if a participant from this condition initially focused on the design of the “oldest” participant currently in the group (e.g. the seventh participant chooses to observe the second) then the rapid replacement rate would mean that the resulting plane was removed from the test group shortly after participant seven themselves began building. Clearly this was not a problem for participants in the one model condition, due to the fact that they only had the

opportunity to observe one other participant (i.e. their immediate predecessor), and the resulting plane would remain on display for the entire duration of their own building period.

The contracted timescale involved in the study is of course one of the great benefits of using micro-society designs. However, in this case, it seems that increasing the time periods involved across the board, for all conditions, might have produced a different result, potentially limiting what we can conclude from the current results. All the same, it remains an open question which of these alternatives would represent a better model of real human cultural learning, given that there will inevitably be certain constraints on how much can be observed by one individual even in large populations.

Our methods offer many possibilities for future research. One particularly interesting avenue for future research concerns the possibility of detecting exactly which social learning strategies (e.g. Laland, 2004) participants are employing. In some of our previous work (Caldwell & Millen, 2008a; 2010) we have taken photographs of the designs produced by individual participants, and this has allowed us to generate similarity scores for particular pairs of designs. Using similar approaches it may be possible to determine whose designs participants are copying, and hence answer questions about the social learning strategies involved. Franz and Nunn (2009; this issue) have also developed methods which allow for investigating social learning strategies in natural contexts, although using a very different approach from our own. In conclusion, micro-society designs have much to offer the study of culture, and can reveal a great deal regarding the longevity of transmitted behaviors, and the ways in which repeated transmission shapes behavior. And this is achieved by (at least up to a point) placing social learning back in context.

Author Note

This work was supported by the Economic and Social Research Council (award numbers RES-061-23-0072 and RES-062-23-1634).

References

- Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, 193, 31-35.
- Baum, W. M., Richerson, P. J., Efferson, C. M. & Paciotti, B. M. (2004). Cultural evolution in laboratory micro-societies including traditions of rule giving and rule following. *Evolution and Human Behavior*, 25, 305-326.
- Bentley, R. A., Hahn, M. W. & Shennan, S. J. (2004). Random drift and culture change. *Proceedings of the Royal Society of London B*, 271, 1443-1450.
- Bentley, R. A., Lipo, C. P., Herzog, H. A. & Hahn, M. W. (2007). Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior*, 28, 151-158.
- Boyd, R. & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.
- Boyd, R. & Richerson, P. J. (1996). Why culture is common but cultural evolution is rare. *Proceedings of the British Academy*, 88, 77-93.
- Caldwell, C. A. & Millen, A. E. (2008a). Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior*, 29, 165-171. doi:10.1016/j.evolhumbehav.2007.12.001

- Caldwell, C. A. & Millen, A. E. (2008b). Studying cumulative cultural evolution in the laboratory. *Philosophical Transactions of the Royal Society B*, 363, 3529–3539. doi:10.1098/rstb.2008.0133
- Caldwell, C. A. & Millen, A. E. (2009). Social learning mechanisms and cumulative cultural evolution: is imitation necessary? *Psychological Science*, 20, 1478–1483. doi:10.1111/J.1467-9280.2009.02469.x
- Caldwell, C. A. & Millen, A. E. (2010). Conservatism in laboratory microsocieties: unpredictable payoffs accentuate group-specific traditions. *Evolution and Human Behavior*, 31, 123-130. doi:10.1016/j.evolhumbehav.2009.08.002
- Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–558.
- Enquist, M., Strimling, P., Eriksson, K., Laland, K. N. & Sjostrand, J. (2010). One cultural parent makes no culture. *Animal Behaviour*. doi:10.1016/j.anbehav.2010.03.009
- Eriksson, K. & Coultas, J. C. (2009). Are people really conformist-biased? An empirical test and a new mathematical model. *Journal of Evolutionary Psychology*, 7, 5–21.
- Franz, M., & Nunn, C. L. (2009). Network-based diffusion analysis: a new method for detecting social learning. *Proceedings of the Royal Society B-Biological Sciences*, 276, 1829-1836.
- Flynn, E. (2008). Investigating children as cultural magnets: do young children transmit redundant information along diffusion chains? *Philosophical Transactions of the Royal Society B*, 363, 3541–3551.
- Henrich, J. (2001). Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominant force in behavioral change. *American Anthropologist*, 103, 992-1013.
- Henrich, J. (2004). Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses – the Tasmanian case. *American Antiquity*, 69, 197-214.
- Henrich, J. & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, 12, 123–135.
- Horner, V. & Whiten, A. (2005) Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Animal Cognition*, 8, 164–181.
- Itani, J. & Nishimura, A. (1973). The study of infrahuman culture in Japan. In E. W. J. Menzel (Ed.) *Precultural Primate Behaviour* (pp. 26-50). Basel: Karger.
- Jacobs, R. C. & Campbell, D. T. (1961). The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *Journal of Abnormal and Social Psychology*, 62, 649-658.
- Kawai, M. (1965). Newly-acquired pre-cultural behavior of the natural troop of Japanese monkeys on Koshima Islet. *Primates*, 6, 1-30.
- Kawamura, S. (1959). The process of sub-culture propagation among Japanese macaques. *Primates*, 2, 43-60.
- Kendal, R. L, Coolen, I., van Bergen, Y. & Laland, K. N. (2005). Trade-offs in the adaptive use of asocial and social learning. *Advances in the Study of Behavior*, 35, 333-379.
- Kirby, S., Cornish, H. & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105, 10681-10686.

- Kirby, S., Dowman, M. & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104, 5241-5245.
- Kirby, S. & Hurford, J. (2002). The emergence of linguistic structure: an overview of the iterated learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language* (pp 121-148). London: Springer Verlag.
- Laland, K. N. (2004). Social learning strategies. *Learning and Behavior*, 32, 4-14.
- Lieberman, E., Michel, J., Jackson, J., Tang, T. & Nowak, M. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449, 713-716.
- Menzel, E. W., Davenport, R. K. & Rogers, C. M. (1972). Protocultural aspects of chimpanzees' responsiveness to novel objects. *Folia Primatologica*, 17, 161-170.
- Mesoudi, A. & Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior*, 30, 41-48.
- Mesoudi, A., Whiten, A. & Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences*, 29, 329-383.
- Page, E. B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58, 216-230.
- Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10, 405-415.
- Pagel, M., Atkinson, Q. D. & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449, 717-720.
- Panger, M. A., Perry, S., Rose, L., Gros-Louis, J., Vogel, E., MacKinnon, K. C. & Baker, M. (2002). Cross-site differences in foraging behavior of white-faced capuchins (*Cebus capucinus*). *American Journal of Physical Anthropology*, 119, 52-66.
- Perry, S., Baker, M., Fedigan, L., Gros-Louis, J., Jack, K., MacKinnon, K. C., Manson, J. H., Panger, M. A., Pyle, K. & Rose, L. (2003). Social conventions in wild white-faced capuchin monkeys: Evidence for traditions in a Neotropical primate. *Current Anthropology*, 44, 241-268.
- Powell, A. Shennan, S. & Thomas, M. G. (2009). Late Pleistocene demography and the appearance of modern human behavior. *Science*, 324, 1298-1301.
- Rogers, E. M. (1995). *Diffusion of innovations*, 4th edition. New York: Free Press.
- Sherif, M. (1936). *The psychology of social norms*. New York: Harper Collins.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.
- Tomasello, M. (2000). Do children have adult syntactic competence? *Cognition*, 74, 209-253.
- Whiten, A., Horner, V., Litchfield, C. & Marshall-Pescini, S. (2004). How do apes ape? *Learning and Behavior*, 32, 36-52.
- Whiten, A. & Mesoudi, A. (2008). Establishing an experimental science of culture: animal social diffusion experiments. *Philosophical Transactions of the Royal Society B*, 363, 3477-3488.